TELECOM Paris

IP PARIS

New methods for efficient Neural Architecture Search

Overview

Deep Neural Networks can solve extremely challenging tasks thanks to complex stacks of (convolutional) layers with thousands of neurons, especially to solve computer vision-related tasks like image classification, object detection or image segmentation. Their success comes from their ability to learn from examples, not requiring any specific expertise and using very general learning strategies, based on loss' minimization. However, many deep models share a common drawback: their growing complexity challenges the computational capability of embedded devices and poses questions around the power consumption when deployed on-the-field.

Neural Architecture Search (NAS) targets automatic research of neural network architectures, optimizing metrics like performance and latency. However, typical NAS approaches are extremely computationally heavy: for example, LEMONADE [1] is an evolutionary algorithm implementing Lamarckism: after every generation, child networks are generated to improve the pareto-frontier with respect to the current population. Other evolutionary algorithms use concepts like monte-carlo optimization [2] or random search [3], which however significantly make the research of optimized architectures extremely difficult and complex to achieve, requiring thousands of computational days to optimize even on smaller datasets.

The main objective of this project is to study some of the most advanced and efficient NAS approaches, identifying improvements on these strategies which are dictated from nature. It is known that most neurons in an infant's brain have relatively few connections to other neurons. During the first two years of life, however, a baby's brain will establish billions of new connections between neurons. The intricacy of neural connections continues to *increase throughout life [4]*. Babies are actually born with many more neurons than they need. In addition, synapses are formed throughout life, based on our changing experiences. Brain development enhances certain capabilities in part by a *pruning down* of unnecessary neurons [4]. As an infant begins to experience the world, neurons that do not become interconnected with other neurons become unnecessary. They eventually die out, increasing the efficiency of the nervous system [4]. Our objective is thereby to model such a behavior, designing a dynamics where in the first stages we deploy a *growing policy,* where neurons are placed and being connected to each other, and a second stage where *pruning* of un-necessary connections is performed. The application of these concepts to NAS covers a central role in this project.

It is well known that many ANNs, trained on some tasks, are typically over-parameterized. The goal of pruning techniques is to achieve the highest sparsity (i.e. the maximum percentage of

removed parameters) with minimal performance loss (accuracy loss versus the “un-pruned” model). Towards this end, a number of different approaches have been proposed: we mention for example the use of variational dropout to promote sparsity [5], the use of sensitivity towards sparse models [6, 7, 8] or the direct minimization of the l0 norm through differentiable proxies [9].
In a recent work, [10] observed that only a certain group of parameters is actually updated during training: this suggests that all the other parameters can be removed from the learning process without affecting the performance. These parameters, however, can be determined a-posteriori only, and other pruning strategies can achieve higher sparsity in general [8]. However, [10] poses some interesting questions: is it possible to learn these sub-structures without training the full model? Recently, the neural growing concept has been proposed [11]; however, it relies on random growing policies and still relies on refined, by-hand hyper-parameter tuning while still not reaching the same performance as other NAS approaches. The main limitation of these policies relies on the fact growing and pruning strategies are each-other independent and simply applied in pipeline.

Objectives at a glance

- Study the most recent state-of-the-art approaches around efficient NAS.
- Understand the underlying learning mechanisms involved in NAS.
- Replicate/reproduce state-of-the-art results, with partial or total re-implementation of the approaches.
- Study the feasibility of efficiency improvements for NAS, inspired from nature and technically borrowed from growing/pruning.

Tutor

Enzo Tartaglione, MdC for the Multimedia equipe enzo.tartaglione@telecom-paris.fr

References

[1] Elsken, T., Metzen, J. H., & Hutter, F. (2018). Efficient multi-objective neural architecture search via lamarckian evolution. *arXiv preprint arXiv:1804.09081*.
[2] Wang, L., Xie, S., Li, T., Fonseca, R., & Tian, Y. (2021). Sample-efficient neural architecture search by learning actions for monte carlo tree search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
[3] Li, L., & Talwalkar, A. (2020, August). Random search and reproducibility for neural architecture search. In *Uncertainty in artificial intelligence* (pp. 367-377). PMLR.
[4] Huttenlocher, P. R. (1979). Synaptic density in human frontal cortex-developmental changes and effects of aging. *Brain Res*, *163*(2), 195-205.
[5] D. Molchanov, A. Ashukha, D. Vetrov, Variational dropout sparsifies deep neural networks, Vol. 5, 2017, pp. 3854–3863.
[6] E. Tartaglione, S. Lepsøy, A. Fiandrotti, G. Francini, Learning sparse neu-ral networks via sensitivity-driven regularization, in: Advances in Neural Information Processing Systems, 2018, pp. 3878–3888.

[7] Tartaglione, E., Bragagnolo, A., Odierna, F., Fiandrotti, A., & Grangetto, M. (2021). SeReNe: Sensitivity-Based Regularization of Neurons for Structured Sparsity in Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*.
[8] Tartaglione E., Bragagnolo A., Grangetto M. (2020) Pruning Artificial Neural Networks: A Way to Find Well-Generalizing, High-Entropy Sharp Minima. In: Farkaš I., Masulli P., Wermter S. (eds) Artificial Neural Networks and Machine Learning – ICANN 2020. ICANN 2020. Lecture Notes in Computer Science, vol 12397. Springer, Cham. https://doi.org/10.1007/978-3-030-61616-8_6
[9] C. Louizos, M. Welling, D. P. Kingma, Learning sparse neural networksthroughl0 regularization, arXiv preprint arXiv:1712.01312.
[10] J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks, 2019.
[11] Li, Y., & Ji, S. (2021, July). Neural plasticity networks. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-9). IEEE.